Lecture 12: Privacy Engineering

CMSC 25910

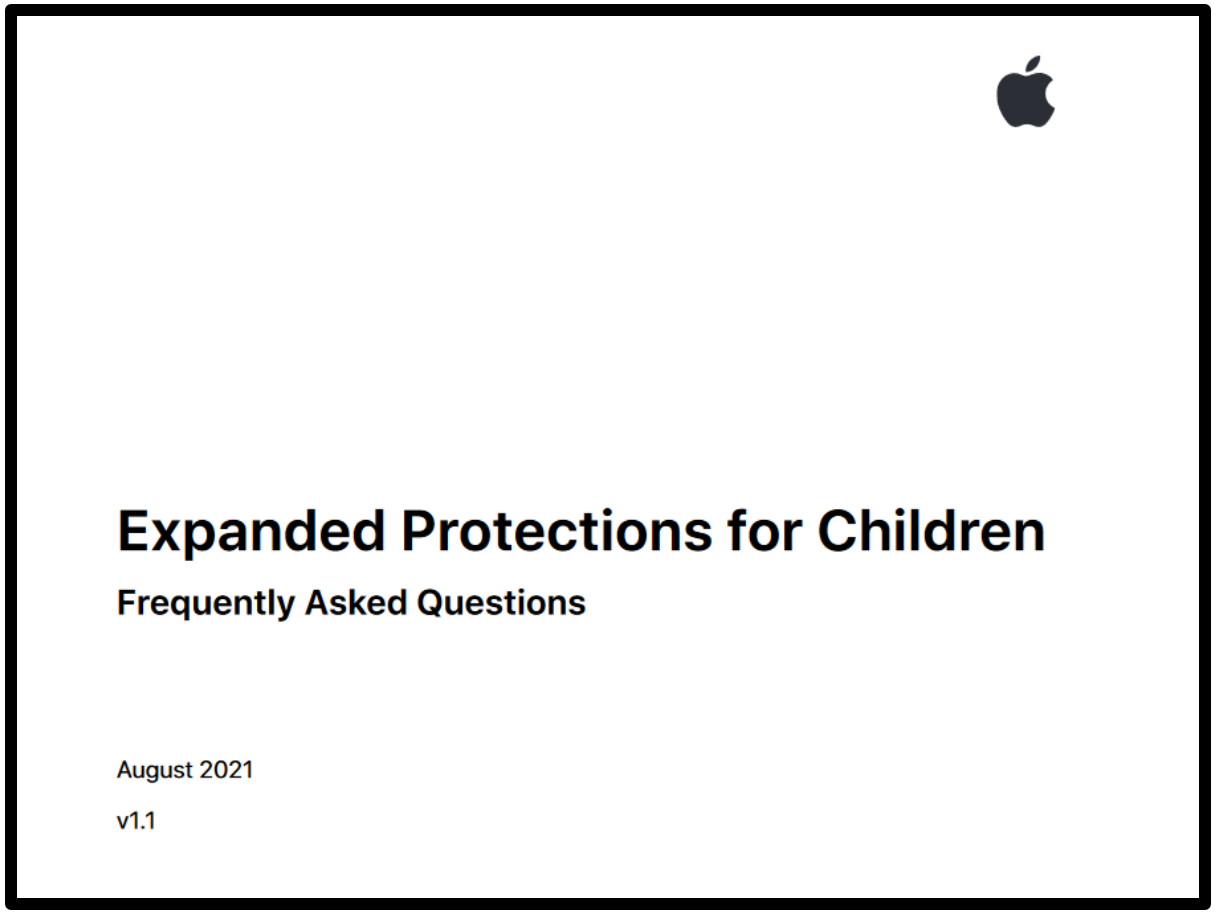
Spring 2024

The University of Chicago

THE UNIVERSITY OF CHICAGO

(A Proposal For) Detecting CSAM in a Privacy-Preserving Way

Apple’s Proposals For CSAM Detection



https://www.apple.com/child-safety/pdf/Expanded_Protections_for_Children_Frequently_Asked_Questions.pdf

Apple’s Proposals For CSAM Detection

- “Apple does not learn anything about images that do not match the known CSAM database.”
- “Apple can’t access metadata or visual derivatives for matched CSAM images until a threshold of matches is exceeded for an iCloud Photos account.”
- “The risk of the system incorrectly flagging an account is extremely low. In addition, Apple manually reviews all reports made to NCMEC to ensure reporting accuracy.”
- “Users can’t access or view the database of known CSAM images.”
- “Users can’t identify which images were flagged as CSAM by the system.”

All points taken verbatim from https://www.apple.com/child-safety/pdf/CSAM_Detection_Technical_Summary.pdf

Apple’s Proposals For CSAM Detection

- “Instead of scanning images in the cloud, the system performs on-device matching using a database of known CSAM image hashes provided by NCMEC and other child-safety organizations. Apple further transforms this database into an unreadable set of hashes, which is securely stored on users’ devices.”

All points taken verbatim from https://www.apple.com/child-safety/pdf/CSAM_Detection_Technical_Summary.pdf

Detour: Hashing

- **Hash functions** are deterministic one-way functions that should be hard to invert (i.e., reverse an output into its corresponding input)

Detour: Hashing

- **Locality-sensitive hashing**: fuzzy hashing that tries to tries to map similar items to the same bucket
  - This often requires you to featurize an image
  - Consider the color distributions, objections recognized, sub-images, and other ways you could partition an image (or other item)
- **Minhash** (min-wise independent permutations locality sensitive hashing scheme) is one type

See https://en.wikipedia.org/wiki/MinHash

Detour: Hashing

- **Perceptual hashing**: a type of locality-sensitive hashing that tries to model how a human would perceive a media object (e.g., picture, audio, video) as its measure of similarity
- Apple's CSAM proposal used Neural Hash, a specific type of perceptual hash function

See https://www.apple.com/child-safety/pdf/CSAM_Detection_Technical_Summary.pdf

Apple's Proposals For CSAM Detection

- **Threshold Secret Sharing:** Cryptographic technique that requires a specified number of shares to reconstruct a secret
- **Private Set Intersection (PSI):** Cryptographic technique that enables two parties to find the intersection of their sets without revealing elements that are not in common

- Broader societal questions: Who controls the database of CSAM? Who decides what objects are included? What happens when possible CSAM is detected?

See https://www.apple.com/child-safety/pdf/CSAM_Detection_Technical_Summary.pdf

Secret Sharing

- (Adi) Shamir's Secret Sharing scheme is based on polynomial interpolation over finite fields
- Insight: k points uniquely determine a polynomial of degree k-1
  - 2 points uniquely define a line, 3 points uniquely define a polynomial of degree 2 ($ax^2 + bx + c$)
- Approach: pick a polynomial of appropriate degree such that the y-intercept is the "secret" and give everyone a point on this polynomial as their "share"

Secure Multiparty Computation (MPC)

- Subfield of cryptography
- Multiple people jointly compute a function on inputs provided by everyone *while keeping those inputs private from each other*
- Adversarial models:
  - Semi-honest / honest-but-curious: Assume participants follow the protocol, but want to learn the private values
  - Malicious: Assume participants may cheat
- Interesting cryptography (you can learn about in other courses)
- Techniques like Private Set Intersection (PSI) are used, for instance, in Apple's password breach monitoring service

Simple MPC Example

- Three people want to compute the average # of succulent plants they have without (shamefully) admitting how many they have
- Each person creates three shares of their own count, distributing them securely to the three people (including themselves)
  - Blase has 150. Makes “shares” 120, -20, 50 (sum to 150)
  - Alison has 20. Makes “shares” 100, -120, 40 (sum to 20)
  - Madison has 10. Makes “shares” 50, -60, 20 (sum to 10)
- Blase ends up with (120, 100, 50); Alison with (-20, -120, -60); Madison with (50, 40, 20).
- Averages: Blase (90); Alison (-66.7); Madison (36.7) = **60**

Private Set Intersection (PSI)

- Enables two parties to find the intersection of sets each hold without learning about the other's items not in the intersection
- Used, for instance, in Apple's password breach monitoring service
  - https://support.apple.com/guide/security/password-monitoring-sec78e79fc3b/web

The underlying protocol partitions the list of curated passwords, which contained approximately 1.5 billion passwords at the time of this writing, into 2^{15} different buckets. The bucket a password belongs to is based on the first 15 bits of the SHA256 hash value of the password. Additionally, each leaked password, pw, is associated with an elliptic curve point on the NIST P256 curve: $P_{pw} = \alpha \cdot H_{SWU}(pw)$, where α is a secret random key known only to Apple and H_{SWU} is a random oracle function that maps passwords to curve points based on the Shallue-van de Woestijne-Ulas method. This transformation is designed to computationally hide the values of passwords and helps prevent revealing newly leaked passwords through Password Monitoring.

To compute the private set intersection, the user's device determines the bucket the user's password belongs to using λ, the 15-bit prefix of SHA256(upw), where upw is one of the user's passwords. The device generates their own random constant, β, and sends the point $P_c = \beta \cdot H_{SWU}(upw)$ to the server, along with a request for the bucket corresponding to λ. Here β hides information about the user's password and limits to λ the information exposed from the password to Apple. Finally, the server takes the point sent by the user's device, computes $\alpha P_c = \alpha\beta \cdot H_{SWU}(upw)$, and returns it, along with the appropriate bucket of points—$B\lambda = \{ P_{pw} \mid \text{SHA256(pw) begins with prefix } \lambda\}$—to the device.

The returned information allows the device to compute $B'_\lambda = \{\beta \cdot P_{pw} \mid P_{pw} \in B_\lambda\}$, and ascertains that the user's password has been leaked if $\alpha P_c \in B'_\lambda$.

Privacy by Design (PbD)

Art. 25 GDPR

Data protection by design and by default

1. Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects.

https://gdpr-info.eu/art-25-gdpr/

Potential Goals of Privacy Engineering

- Compliance with laws
  - GDPR, CCPA, etc.
- Compliance with reasonable consumer expectations and making accurate public statements
  - See, e.g., FTC's mandates
- Engender trust and goodwill among your users
  - "Competing" based on privacy-protectiveness
- Protecting privacy as a societal value / "it's the right thing"

Mechanisms

- Thinking carefully and being selective about data collection
  - Data minimization, immediate de-identification/pseudonymization
- Not retaining data
- Thoughtful applications of cryptographic tools
- Thoughtful architectures for computer systems
  - Access control (locked down? open with audit logs?), data storage
- Public statements and user communication (e.g., in UIs)
- Appropriate socio-technical processes, audits, and reviews

Privacy Impact Assessments (PIAs)

- "A PIA is an **analysis of how personally identifiable information is collected, used, disseminated, and maintained**. It examines how the Department has **incorporated privacy concerns throughout its development, design, and deployment** of a technology, program, or rulemaking. "Personally identifiable information" is defined as any information that permits the identity of an individual to be directly or indirectly inferred, including any other information which is linked or linkable to that individual"

https://www.dhs.gov/sites/default/files/publications/privacy_pia_guidance_june2010_0.pdf

Privacy Impact Assessments (PIAs)

- “The purpose of a PIA is to demonstrate that program managers and system owners have **consciously incorporated privacy protections throughout the development life cycle** of a system or program. This involves making certain that privacy protections are **built into the system from the initiation of development**, not after the fact when they can be far more costly or could affect the viability of the project. The PIA process requires that candid and forthcoming communications occur between the program manager, system owner, the component’s Privacy Officer, and the Privacy Office to ensure appropriate and timely handling of privacy concerns. Addressing privacy issues publicly through a PIA builds citizen trust...”

https://www.dhs.gov/sites/default/files/publications/privacy_pia_guidance_june2010_0.pdf

Case Studies

1: In-store Tracking

- **Setting:** BlaseMart wants to track customers as they move through the store to:
  - Understand which areas of the store are most popular
  - Optimize the relative location of different displays by understanding which sections are highly correlated among consumers' visits
  - Provide targeted discounts to specific consumers about specific products

2: Maps App

- **Setting:** Blaze (like Waze, but better) wants to provide a privacy-protective alternative to Google Maps
  - Real-time traffic info
  - Accident/closure reporting
  - Convenient history-based recommendations for users

3: Browser

- **Setting:** Blaze Browser (like Brave, but better) wants to determine what malware websites its users have visited
  - Caveat: The malware sites are only identified after the fact

4: Voice Assistant

- **Setting:** Blazezon wants to improve the speech recognition of Alexa on the Blazezon Echo

Learning More About Privacy

Key Privacy Organizations / Nonprofits

- International Association of Privacy Professionals (IAPP)
  - https://iapp.org
  - Provide certifications like CIPM, CIPP/US, and CIPT
- Future of Privacy Forum (FPF)
  - https://fpf.org/
  - Think tank and advocacy group
- Electronic Frontier Foundation (EFF)
  - https://eff.org/
  - Nonprofit advocacy group defending digital privacy

Privacy Engineering Practice & Respect

- USENIX Conference on Privacy Engineering & Respect (PEPR)
- https://usenix.org/conference/pepr24